

Nested Quantum Annealing Correction

Walter Vinci,^{1,2,3} Tameem Albash,^{2,3,4} and Daniel A. Lidar^{1,2,3,5}

¹*Department of Electrical Engineering, University of Southern California, Los Angeles, California 90089, USA*

²*Department of Physics and Astronomy, University of Southern California, Los Angeles, California 90089, USA*

³*Center for Quantum Information Science & Technology,*

University of Southern California, Los Angeles, California 90089, USA

⁴*Information Sciences Institute, University of Southern California, Marina del Rey, California 90292, USA*

⁵*Department of Chemistry, University of Southern California, Los Angeles, California 90089, USA*

We present a general error-correcting scheme for quantum annealing that allows for the encoding of a logical qubit into an arbitrarily large number of physical qubits. Given any Ising model optimization problem, the encoding replaces each logical qubit by a complete graph of degree C , representing the distance of the error-correcting code. A subsequent minor-embedding step then implements the encoding on the underlying hardware graph of the quantum annealer. We demonstrate experimentally that the performance of a D-Wave Two quantum annealing device improves as C grows. We show that the performance improvement can be interpreted as arising from an effective increase in the energy scale of the problem Hamiltonian, or equivalently, an effective reduction in the temperature at which the device operates. The number C thus allows us to control the amount of protection against thermal and control errors, and in particular, to trade qubits for a lower effective temperature that scales as $C^{-\eta}$, with $\eta \leq 2$. This effective temperature reduction is an important step towards scalable quantum annealing.

Quantum annealing (QA) attempts to exploit quantum fluctuations to solve computational problems faster than it is possible with classical computers [1–7]. As an approach designed to solve optimization problems, QA is a special case of adiabatic quantum computation (AQC) [8], a universal model of quantum computing [9–12]. In AQC, a system is designed to follow the instantaneous ground state of a time-dependent Hamiltonian whose final ground state encodes the solution to the problem of interest. This results in a certain amount of stability, since the system can thermally relax to the ground state after an error, as well as resilience to errors, since the presence of a finite energy gap suppresses thermal and dynamical excitations [13–18].

Despite this inherent robustness to certain forms of noise, AQC requires error-correction to ensure scalability, just like any other form of quantum information processing [19]. Various error correction proposals for AQC and QA have been made [20–33], but an accuracy-threshold theorem for AQC is not yet known, unlike in the circuit model (e.g., [34]). A direct AQC simulation of a fault-tolerant quantum circuit leads to many-body (high-weight) operators that are difficult to implement [23, 24] or myriad other problems [12]. Nevertheless, a scalable method to reduce the effective temperature would go a long way towards approaching the ideal of closed-system AQC, where only non-adiabatic transitions constitute the source of errors.

Motivated by the availability of commercial QA devices featuring hundreds of qubits [35–38], we focus on error correction for QA. There is a consensus that these devices are significantly and adversely affected by decoherence, noise, and control errors [39–46], which makes them particularly interesting for the study of tailored, practical error correction techniques. Such techniques, known as quantum annealing correction (QAC) schemes, have already been experimentally shown to significantly improve the performance of quantum annealers [26, 30–32], and theoretically analyzed using a mean-field approach [33]. However, these QAC schemes are not easily

generalizable to arbitrary optimization problems since they induce an encoded graph that is typically of a lower degree than the qubit-connectivity graph of the physical device. Moreover, they typically impose a fixed code distance, which limits their efficacy.

To overcome these limitations, here we present a family of error-correcting codes for QA, based on a “nesting” scheme, that has the following properties: (1) it can handle arbitrary Ising-model optimization problem, (2) it can be implemented on present-day QA hardware, and (3) it is capable of an effective temperature reduction controlled by the code distance. Our “nested quantum annealing correction” (NQAC) scheme thus provides a very general and practical tool for error correction in quantum optimization.

We test NQAC by studying antiferromagnetic complete graphs numerically, as well as on a D-Wave Two (DW2) processor featuring 504 flux qubits connected by 1427 tunable composite qubits acting as Ising-interaction couplings, arranged in a non-planar Chimera-graph lattice [47] (complete graphs were also studied for a spin glass model in Ref. [48]). We demonstrate that our encoding schemes yields a steady improvement for the probability of reaching the ground state as a function of the nesting level, even after minor-embedding the complete graph onto the physical graph of the quantum annealer. We also demonstrate that NQAC outperforms classical repetition code schemes that use the same number of physical qubits.

I. QUANTUM ANNEALING AND ENCODING THE HAMILTONIAN

In QA the system undergoes an evolution governed by the following time-dependent, transverse-field Ising Hamiltonian:

$$H(t) = A(t)H_X + B(t)H_P, \quad t \in [0, t_f], \quad (1)$$

with respectively monotonically decreasing and increasing “annealing schedules” $A(t)$ and $B(t)$. The “driver Hamiltonian” $H_X = -\sum_i \sigma_i^x$ is a transverse field whose amplitude controls the tunneling rate. The solution to an optimization problem of interest is encoded in the ground state of the Ising problem Hamiltonian H_P , with

$$H_P = \sum_{i \in \mathcal{V}} h_i \sigma_i^z + \sum_{(i,j) \in \mathcal{E}} J_{ij} \sigma_i^z \sigma_j^z, \quad (2)$$

where the sums run over the weighted vertices \mathcal{V} and edges \mathcal{E} of a graph $G = (\mathcal{V}, \mathcal{E})$, and $\sigma_i^{x,z}$ denote the Pauli operators acting on qubit i . The D-Wave devices use an array of superconducting flux qubits to physically realize the system described in Eqs. (1) and (2) on a fixed “Chimera” graph (see Fig. 1) with programmable local fields $\{h_i\}$, couplings $\{J_{ij}\}$, and annealing time t_f [36–38].

For closed systems, the adiabatic theorem [49, 50] guarantees that if the system is initialized in the ground state of $H(0) = A(0)H_X$, a sufficiently slow evolution relative to the inverse minimum gap of $H(t)$ will take the system with high probability to the ground state of the final Hamiltonian $H(t_f) = B(t_f)H_P$. Dynamical errors then arise due to diabatic transitions, but they can be made arbitrarily small via boundary cancellation methods that control the smoothness of $A(t)$ and $B(t)$, as long as the adiabatic condition is satisfied [51–53]. For open systems, specifically a system that is weakly coupled to a thermal environment, the final state is a mixed state $\rho(t_f)$ that is close to the Gibbs state associated with $H(t_f)$ if equilibration is reached throughout the annealing process [18, 54, 55]. In the adiabatic limit the open system QA process is thus better viewed as a Gibbs distribution sampler. The main goal of QAC is to suppress the associated thermal errors and restore the ability of QA to act as a ground state solver. In addition QAC should suppress errors due to noise-driven deviations in the specification of H_P [25].

Error correction is achieved in QAC by mapping the logical Hamiltonian $H(t)$ to an appropriately chosen encoded Hamiltonian $\tilde{H}(t)$:

$$\tilde{H}(t) = A(t)H_X + B(t)\tilde{H}_P, \quad t \in [0, t_f], \quad (3)$$

defined over a set of physical qubits \bar{N} larger than the number of logical qubits $N = |\mathcal{V}|$. Note that \tilde{H}_P also includes penalty terms, as explained below. The logical ground state of H_P is extracted from the encoded system’s state $\bar{\rho}(t_f)$ through an appropriate decoding procedure. A successful error correction scheme should recover the logical ground state with a higher probability than a direct implementation of H_P , or than a classical repetition code using the same number of physical qubits \bar{N} . Due to practical limitations of current QA devices that prevent the encoding of H_X , only H_P is encoded in QAC.

In order to allow for the most general N -variable Ising optimization problem, we now define an encoding procedure for problem Hamiltonians H_P supported on a complete graph K_N . The first step of our construction involves a “nested” Hamiltonian \tilde{H}_P that is defined by embedding the logical K_N into a larger $K_{C \times N}$. The integer C is the “nesting level” and controls the amount of hardware resources (qubits, couplers,

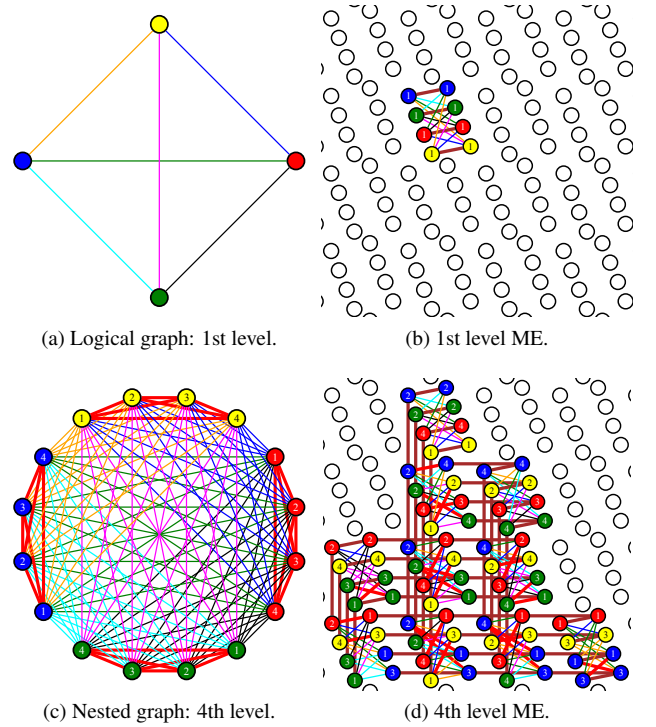


FIG. 1. Illustration of the nesting scheme. In the left column, a C -level nested graph is constructed by embedding a K_N into a $K_{C \times N}$, with $N = 4$ and $C = 1$ (top) and $C = 4$ (bottom). Red, thick couplers are energy penalties defined on the nested graph between the (i, c) nested copies of each logical qubit i . The right column shows the nested graphs after ME on the DW2 Chimera graph. Brown, thick couplers correspond to the ferromagnetic chains introduced in the process.

and local fields) used to represent the logical problem. \tilde{H}_P is constructed as follows. Each logical qubit i ($i = 1, \dots, N$) is represented by a C -tuple of encoded qubits (i, c) , with $c = 1, \dots, C$. The “nested” couplers $\tilde{J}_{(i,c),(j,c')}$ and local fields $\tilde{h}_{(i,c)}$ are then defined as follows:

$$\tilde{J}_{(i,c),(j,c')} = J_{ij}, \quad \forall c, c', i \neq j, \quad (4a)$$

$$\tilde{h}_{(i,c)} = Ch_i, \quad \forall c, i, \quad (4b)$$

$$\tilde{J}_{(i,c),(i,c')} = -\gamma, \quad \forall c \neq c'. \quad (4c)$$

This construction is illustrated in the left column of Fig. 1. Each logical coupling J_{ij} has C^2 copies $\tilde{J}_{(i,c),(j,c')}$, thus boosting the energy scale at the encoded level by a factor of C^2 . Each local field h_i has C copies $\tilde{h}_{(i,c)}$; the factor C in Eq. (4b) ensures that the energy boost is equalized with the couplers. For each logical qubit i , there are $C(C-1)/2$ ferromagnetic couplings $\tilde{J}_{(i,c),(i,c')}$ of strength $\gamma > 0$ (to be optimized), representing energy penalties that promote agreement among the C encoded qubits, i.e., that bind the C -tuple as a single logical qubit i .

The second step of our construction is to implement the fully connected problem \tilde{H}_P on given QA hardware, with a

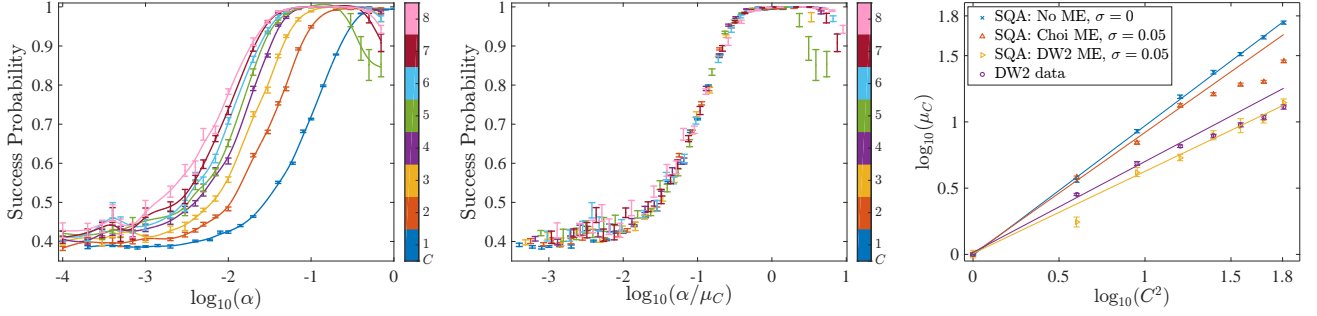


FIG. 2. Experimental and numerical results for the antiferromagnetic K_4 , after encoding, followed by ME and decoding. Left: DW2 success probabilities $P_C(\alpha)$ for eight nesting levels C . Increasing C generally increases $P_C(\alpha)$ at fixed α . Middle: Rescaled $P_C(\alpha/\mu_C)$ data, exhibiting data-collapse. Right: scaling of the energy boost μ_C vs the maximal energy boost μ_C^{\max} , for both the DW2 and SQA. Purple circles: DW2 results. Blue stars: SQA for the case of no ME (i.e., for the problem defined directly over $K_{C \times N}$ and no coupler noise). Red up-triangles: SQA for the Choi ME [56] (for a full Chimera graph), with $\sigma = 0.05$ Gaussian noise on the couplings. Yellow right-triangles: SQA for the DW2 heuristic ME [57, 58] (applied to a Chimera graph with 8 missing qubits) with $\sigma = 0.05$ Gaussian noise on the couplings. The flattening of μ_C suggests that the energy boost becomes less effective at large C . However, this can be remedied by increasing the number of SQA sweeps (see Appendix C), fixed here at 10^4 . Thus the lines represent best fits to only the first four data points, with slopes 0.98, 0.91, 0.62 and 0.69 respectively. In all panels $N_{\text{phys}} \in [8, 288]$.

lower-degree qubit connectivity graph. This requires a minor embedding (ME) [56–60]. The procedure involves replacing each qubit in \tilde{H}_P by a ferromagnetically-coupled chain of qubits, such that all couplings in \tilde{H}_P are represented by inter-chain couplings. The intra-chain coupling represents another energy penalty that forces the chain qubits to behave as a single logical qubit. The physical Hamiltonian obtained after this ME step is the final encoded Hamiltonian \bar{H}_P . We can minor-embed a $K_{C \times N}$ nested graph representing each qubit (i, c) as a physical chain of length $L = \lceil CN/4 \rceil + 1$ on the Chimera graph [56]. This is illustrated in the right column of Fig. 1. The number of physical qubits necessary for a ME of a $K_{C \times N}$ is $N_C^{\text{phys}} = CNL \sim C^2 N^2/4$.

At the end of a QA run implementing the encoded Hamiltonian \bar{H}_P and a measurement of the physical qubits, a decoding procedure must be employed to recover the logical state. For the sake of simplicity we only consider majority vote decoding over both the length- L chain of each encoded qubit (i, c) and the C encoded qubits comprising each logical qubit i (decoding over the length- L chain first, then over the C encoded qubits, does not affect performance; see Appendix A). The encoded and logical qubits can thus be viewed as forming repetition codes with, respectively, distance L and C . Other decoding strategies are possible wherein the encoded or logical qubits do not have this simple interpretation; e.g., energy minimization decoding, which tends to outperform majority voting [31]. In the unlikely event of a tie, we assign a random value of $+1$ or -1 to the logical qubit.

II. RESULTS

Free energy – Using a mean-field analysis similar to the approach pursued in Ref. [33] we can compute the partition

function associated with the nested Hamiltonian $A(t)H_X + B(t)\tilde{H}_P$ for the case with uniform antiferromagnetic couplings. This leads to the following free energy density in the low temperature and thermodynamic limits (see Appendix B):

$$\beta F = C^2 \beta \left(\sqrt{[A(t)/C]^2 + [2\gamma B(t)m]^2} - \gamma B(t)m^2 \right) \quad (5)$$

where m is the mean-field magnetization. There are two key noteworthy aspects of this result. First, the driver term is rescaled as $A(t) \mapsto C^{-1}A(t)$. This shifts the crossing between the A and B annealing schedules to an earlier point in the evolution and is related to the fact that QAC encodes only the problem Hamiltonian term proportional to $B(t)$. Consequently the quantum critical point is moved to earlier in the evolution, which benefits QAC since the effective energy scale at this new point is higher [33]. Second, the inverse temperature is rescaled as $\beta \mapsto C^2 \beta$. This corresponds to an effective temperature reduction by C^2 , a manifestly beneficial effect. The same conclusion, of a lower effective temperature, is reached by studying the numerically computed success probability associated with thermal distributions (see Appendix C). We shall demonstrate that this prediction is born out by our experimental results, though it is masked to some extent by complications arising from the ME and noise.

NQAC results – The hardness of an Ising optimization problem, using a QA device, is controlled by its size N as well as by an overall energy scale α [42]. The smaller this energy scale, the higher the effective temperature and the more susceptible QA becomes to (dynamical and thermal) excitations out of the ground state and misspecification noise on the problem Hamiltonian. This provides us with an opportunity to test NQAC. Since in our experiments we were limited by the largest complete graph that can be embedded on the DW2 device, a K_{32} (see Appendix D for details), we tuned the hardness of a problem by studying the performance of NQAC as

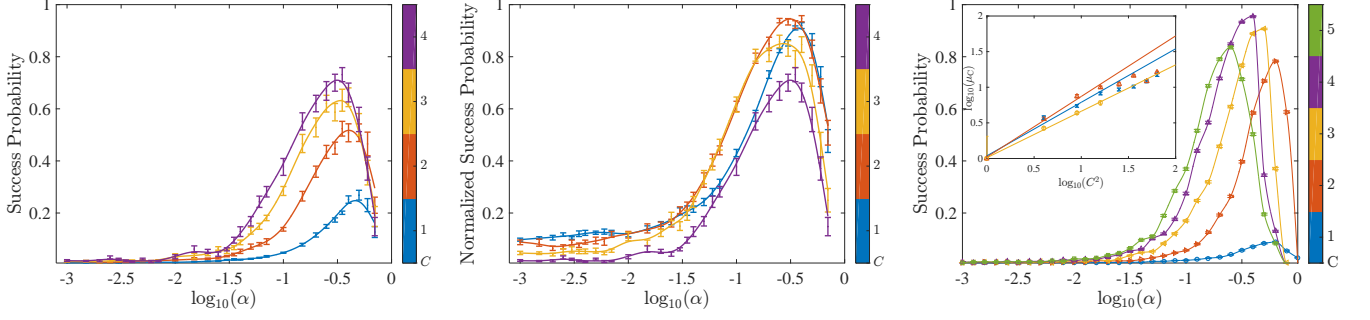


FIG. 3. Random antiferromagnetic K_8 : experimental and numerical results. Left: success probabilities $P_C(\alpha)$ for four nesting levels. Middle: success probabilities $P'_C(\alpha)$ adjusted for classical repetition. Right: numerical results for SQA simulations with 20000 sweeps, $\sigma = 0.05$ Gaussian noise on the couplings, and with the Choi embedding, showing five nesting levels. Inset: scaling of the energy boost μ_C vs the maximal energy boost μ_C^{\max} , for both the DW2 and SQA. Yellow circles: DW2 results. Blue crosses and red up-triangles: SQA for the Choi ME with 10000 (crosses) and 20000 (up-triangles) sweeps, and with $\sigma = 0.05$ Gaussian noise on the couplings. The flattening of μ_C for $C > 4$ suggests that the energy boost becomes less effective at large C , but increasing the number of sweeps recovers the effectiveness. The lines represent best fits to only the first four data points, with respective slopes $\eta/2 = 0.65, 0.75$, and 0.85 .

a function of α via $H_P \mapsto \alpha H_P$, with $0 < \alpha \leq 1$. Note that we did not rescale γ ; instead γ was optimized for optimal post-decoding performance (see Appendix E). It is known that for the DW2, intrinsic coupler control noise can be taken to be Gaussian with standard deviation $\sigma \sim 0.05$ of the maximum value for the couplings [45]. Thus we may expect that, without error correction, Ising problems with $\alpha \lesssim 0.05$ are dominated by control noise.

We applied NQAC to completely antiferromagnetic ($h_i = 0 \forall i$) Ising problems over K_4 ($J_{ij} = 1 \forall i, j$), and K_8 (random $J_{ij} \in [0.1, 1]$ with steps of 0.1) with nesting up to $C = 8$ and $C = 4$, respectively. We denote by $P_C(\alpha)$ the probability to obtain the logical ground state at energy scale α for the C -level nested implementation (see Appendix A for data collection methods). The experimental QA data in Fig. 2 (left) shows a monotonic increase of $P_C(\alpha)$ as a function of the nesting level C over a wide range of energy scales α . As expected, $P_C(\alpha)$ drops from $P_C(1) = 1$ (solution always found) to $P_C(0) = 6/16$ (random sampling of 6 ground states, where 4 out of the 6 couplings are satisfied, out of a total of 16 states).

Note that $P_1(\alpha)$ (no nesting) drops by $\sim 50\%$ when $\alpha \sim 0.1$, which is consistent with the aforementioned $\sigma \sim 0.05$ control noise level, while $P_8(\alpha)$ exhibits a similar drop only when $\alpha \sim 0.01$. This suggests that NQAC is particularly effective in mitigating the two dominant effects that limit the performance of quantum annealers: thermal excitations and control errors. To investigate this more closely, the middle panel of Fig. 2 shows that the data from the left panel can be collapsed via $P_C(\alpha) \mapsto P_C(\alpha/\mu_C)$, where μ_C is an empirical rescaling factor discussed below (see also Appendix F). This implies that $P_1(\mu_C \alpha) \approx P_C(\alpha)$, and hence that the performance enhancement obtained at nesting level C can be interpreted as an energy boost $\alpha \mapsto \mu_C \alpha$ with respect to an implementation without nesting.

The existence of this energy boost is a key feature of NQAC, as anticipated above. Recall [Eq. (4)] that a nested

graph $K_{C \times N}$ contains C^2 equivalent copies of the same logical coupling J_{ij} . Hence a level- C nesting before ME can provide a maximal energy boost $\mu_C^{\max} = C^{\eta^{\max}}$, with $\eta^{\max} = 2$. This simple argument agrees with the reduction of the effective temperature by C^2 based on the calculation of the free energy (5). The right panel of Fig. 2 shows μ_C as a function of μ_C^{\max} , yielding $\mu_C \sim C^\eta$ with $\eta \approx 1.37$ (purple circles). To understand why $\eta < \eta^{\max}$, we performed simulated quantum annealing (SQA) simulations (see Appendix G for details). We observe in Fig. 2 (right) that without ME and control errors, the boost scaling matches μ_C^{\max} (blue stars). When including ME and control errors a performance drop results (red triangles). Both factors thus contribute to the sub-optimal energy boost observed experimentally. However, the optimal energy boost is recovered for a fully thermalized state with a sufficiently large penalty (see Appendix C). To match the experimental DW2 results using SQA we replace the Choi ME designed for full Chimera graphs [56] by the heuristic ME designed for Chimera graphs with missing qubits [57, 58], and achieve a near match (yellow triangles) (see Appendix D for more details on ME).

Performance of NQAC vs classical repetition – Recall that $N_C^{\text{phys}} = CNL$ is the total number of physical qubits used at nesting level C ; let C_{\max} denote the highest nesting level that can be accommodated on the QA device for a given K_N , i.e., $C_{\max}NL \leq N_{\text{tot}} < (C_{\max} + 1)NL$, where N_{tot} is the total number of physical qubits (504 in our experiments). Then $M_C = \lfloor N_C^{\text{phys}} / N_C^{\text{phys}} \rfloor$ is the number of copies that can be implemented in parallel. For NQAC at level C to be useful, it must be more effective than a classical repetition scheme where M_C copies of the problem are implemented in parallel. If a single implementation has success probability $P_C(\alpha)$, the probability to succeed at least once with M_C statistically independent implementations is $P'_C(\alpha) = 1 - [1 - P_C(\alpha)]^{M_C}$. It turns out that the antiferromagnetic K_4 problem, for which a random guess succeeds

with probability $6/16$, is too easy [i.e., $P'_C(\alpha)$ approaches 1 too rapidly], and we therefore consider a harder problem: an antiferromagnetic K_8 instance with couplings randomly generated from the set $J_{ij} \in \{0.1, 0.2, \dots, 0.9, 1\}$ (see Appendix E for more details and data on this and additional instances). Problems of this type turn out to have a sufficiently low success probability for our purposes, and can still be nested up to $C = 4$ on the DW2 processor.

Results for $P_C(\alpha)$ are shown in Fig. 3 (left), and again increase monotonically with C , as in the K_4 case. For each C , $P_C(\alpha)$ peaks at a value of α for which the maximum allowed strength of the energy penalties $\gamma = 1$ is optimal ($\gamma > 1$ would be optimal for larger α , as shown in Appendix E; the growth of the optimal penalty with problem size, and hence chain length, is a typical feature of minor-embedded problems [48]). An energy-boost interpretation of the experimental data of Fig. 3 is possible for α values to the left of the peak; to the right of the peak, the performance is hindered by the saturation of the energy penalties.

Figure 3 (middle) compares the success probabilities $P'_C(\alpha)$ adjusted for classical repetition, where we have set $C_{\max} = 4$, and shows that $P'_2(\alpha) > P'_1(\alpha)$, i.e., even after accounting for classical parallelism $C = 2$ performs better than $C = 1$. However, we also find that $P'_4(\alpha) < P'_3(\alpha) \leq P'_2(\alpha)$, so no additional gain results from increasing C in our experiments. This can be attributed to the fact that even the K_8 problem still has a relatively large $P_1(\alpha)$. Experimental tests on QA devices with more qubits will thus be important to test the efficacy of higher nesting levels on harder problems.

To test the effect of increasing C , and also to study the effect of varying the annealing time, we present in Fig. 3 (right) the performance of SQA on a random K_8 antiferromagnetic instance with the Choi ME. The results are qualitatively similar to those observed on the DW2 processor with the heuristic ME [Fig. 3 (left)]. Interestingly, we observe a drop in the peak performance at $C = 5$ relative to the peak observed for $C = 4$. We attribute this to both a saturation of the energy penalties and a suboptimal number of sweeps. The latter is confirmed in Fig. 3 (right, inset), where we observe that the scaling of μ_C with C is better for the case with more sweeps, i.e., again $\mu_C \sim C^\eta$, and η increases with the number of sweeps.

III. DISCUSSION

Nested QAC offers several significant improvements over previous approaches to the problem of error correction for QA. It is a flexible method that can be used with any optimization problem, and allows the construction of a family of codes with arbitrary code distance. We have given experimental and numerical evidence that nesting is effective by performing studies with a D-Wave QA device and numerical simulations. We have demonstrated that the protection from errors provided by NQAC can be interpreted as arising from an increase (with nesting level C) in the energy scale at which the logical problem is implemented. This represents a very useful tradeoff: the effective temperature drops as we increase the number of qubits allocated to the encoding, so that these two resources can be traded. Thus NQAC can be used to combat thermal excitations, which are the dominant source of errors in QA, and are the bottleneck for scalable QA implementations. We have also demonstrated that an appropriate nesting level can outperform classical repetition with the same number of qubits, with improvements to be expected when next-generation QA devices with larger numbers of physical qubits become available. We, therefore, believe that our results are of immediate and near-future practical use, and constitute an important step toward scalable QA.

ACKNOWLEDGEMENTS

We thank Prof. Hidetoshi Nishimori and Dr. Shunji Matsuura for valuable comments, and Dr. Aidan Roy for providing the minor embeddings used in the experiments with the D-Wave Two. Access to the D-Wave Two was made available by the USC-Lockheed Martin Quantum Computing Center. Part of the computing resources were provided by the USC Center for High Performance Computing and Communications. This work was supported under ARO grant number W911NF-12-1-0523, ARO MURI Grant Nos. W911NF-11-1-0268 and W911NF-15-1-0582, and NSF grant number INSPiRE-1551064.

[1] Kadowaki, T. & Nishimori, H. Quantum annealing in the transverse Ising model. *Phys. Rev. E* **58**, 5355 (1998).
[2] Brooke, J., Bitko, D., F., T., Rosenbaum & Aeppli, G. Quantum Annealing of a Disordered Magnet. *Science* **284**, 779–781 (1999). URL <http://www.sciencemag.org/content/284/5415/779.abstract>.
[3] Brooke, J., Rosenbaum, T. F. & Aeppli, G. Tunable quantum tunnelling of magnetic domain walls. *Nature* **413**, 610–613 (2001).

[4] Farhi, E. *et al.* A Quantum Adiabatic Evolution Algorithm Applied to Random Instances of an NP-Complete Problem. *Science* **292**, 472–475 (2001). URL <http://www.sciencemag.org/content/292/5516/472>.
[5] Morita, S. & Nishimori, H. Mathematical foundation of quantum annealing. *J. Math. Phys.* **49**, 125210–47 (2008).
[6] Das, A. & Chakrabarti, B. K. *Colloquium: Quantum annealing and analog quantum computation*. *Rev. Mod. Phys.* **80**, 1061–1081 (2008).

- [7] S. Suzuki and A. Das (guest eds.). Discussion and Debate - Quantum Annealing: The Fastest Route to Quantum Computation? *Eur. Phys. J. Spec. Top.* **224**, 1 (2015). URL <http://epjst.epj.org/articles/epjst/abs/2015/01/contents/contents.html>.
- [8] Farhi, E., Goldstone, J., Gutmann, S. & Sipser, M. Quantum Computation by Adiabatic Evolution. *arXiv:quant-ph/0001106* (2000). URL <http://arxiv.org/abs/quant-ph/0001106>.
- [9] Aharonov, D. *et al.* Adiabatic Quantum Computation is Equivalent to Standard Quantum Computation. *SIAM J. Comput.* **37**, 166–194 (2007).
- [10] Mizel, A., Lidar, D. A. & Mitchell, M. Simple Proof of Equivalence between Adiabatic Quantum Computation and the Circuit Model. *Phys. Rev. Lett.* **99**, 070502 (2007). URL <http://link.aps.org/doi/10.1103/PhysRevLett.99.070502>.
- [11] Gosset, D., Terhal, B. M. & Vershynina, A. Universal Adiabatic Quantum Computation via the Space-Time Circuit-to-Hamiltonian Construction. *Phys. Rev. Lett.* **114**, 140501– (2015). URL <http://link.aps.org/doi/10.1103/PhysRevLett.114.140501>.
- [12] Lloyd, S. & Terhal, B. Adiabatic and Hamiltonian computing on a 2D lattice with simple 2-qubit interactions. *arXiv:1509.01278* (2015). URL <http://arXiv.org/abs/1509.01278>.
- [13] Childs, A. M., Farhi, E. & Preskill, J. Robustness of adiabatic quantum computation. *Phys. Rev. A* **65**, 012322 (2001).
- [14] Sarandy, M. S. & Lidar, D. A. Adiabatic Quantum Computation in Open Systems. *Phys. Rev. Lett.* **95**, 250503– (2005). URL <http://link.aps.org/doi/10.1103/PhysRevLett.95.250503>.
- [15] Amin, M. H. S., Love, P. J. & Truncik, C. J. S. Thermally Assisted Adiabatic Quantum Computation. *Phys. Rev. Lett.* **100**, 060503 (2008).
- [16] Lloyd, S. Robustness of Adiabatic Quantum Computing. *arXiv:0805.2757* (2008). URL <http://arXiv.org/abs/0805.2757>.
- [17] Amin, M. H. S., Averin, D. V. & Nesteroff, J. A. Decoherence in adiabatic quantum computation. *Phys. Rev. A* **79**, 022107 (2009). URL <http://link.aps.org/doi/10.1103/PhysRevA.79.022107>.
- [18] Albash, T. & Lidar, D. A. Decoherence in adiabatic quantum computation. *Phys. Rev. A* **91**, 062320– (2015). URL <http://link.aps.org/doi/10.1103/PhysRevA.91.062320>.
- [19] Lidar, D. & Brun, T. (eds.). *Quantum Error Correction* (Cambridge University Press, Cambridge, UK, 2013). URL <http://www.cambridge.org/9780521897877>.
- [20] Jordan, S. P., Farhi, E. & Shor, P. W. Error-correcting codes for adiabatic quantum computation. *Phys. Rev. A* **74**, 052322 (2006). URL <http://link.aps.org/doi/10.1103/PhysRevA.74.052322>.
- [21] Lidar, D. A. Towards Fault Tolerant Adiabatic Quantum Computation. *Phys. Rev. Lett.* **100**, 160506 (2008). URL <http://link.aps.org/doi/10.1103/PhysRevLett.100.160506>.
- [22] Quiroz, G. & Lidar, D. A. High-fidelity adiabatic quantum computation via dynamical decoupling. *Phys. Rev. A* **86**, 042333 (2012).
- [23] Young, K. C., Sarovar, M. & Blume-Kohout, R. Error Suppression and Error Correction in Adiabatic Quantum Computation: Techniques and Challenges. *Phys. Rev. X* **3**, 041013– (2013). URL <http://link.aps.org/doi/10.1103/PhysRevX.3.041013>.
- [24] Sarovar, M. & Young, K. C. Error suppression and error correction in adiabatic quantum computation: non-equilibrium dynamics. *New J. of Phys.* **15**, 125032 (2013). URL <http://stacks.iop.org/1367-2630/15/i=12/a=125032>.
- [25] Young, K. C., Blume-Kohout, R. & Lidar, D. A. Adiabatic quantum optimization with the wrong Hamiltonian. *Phys. Rev. A* **88**, 062314– (2013). URL <http://link.aps.org/doi/10.1103/PhysRevA.88.062314>.
- [26] Pudenz, K. L., Albash, T. & Lidar, D. A. Error-corrected quantum annealing with hundreds of qubits. *Nat. Commun.* **5**, 3243 (2014). URL dx.doi.org/10.1038/ncomms4243.
- [27] Ganti, A., Onunkwo, U. & Young, K. Family of $[[6k, 2k, 2]]$ codes for practical, scalable adiabatic quantum computation. *Phys. Rev. A* **89**, 042313– (2014). URL <http://link.aps.org/doi/10.1103/PhysRevA.89.042313>.
- [28] Bookatz, A. D., Farhi, E. & Zhou, L. Error suppression in Hamiltonian-based quantum computation using energy penalties. *Physical Review A* **92**, 022317– (2015). URL <http://link.aps.org/doi/10.1103/PhysRevA.92.022317>.
- [29] Mizel, A. Fault-tolerant, Universal Adiabatic Quantum Computation. *arXiv:1403.7694* (2014). URL <http://arXiv.org/abs/1403.7694>.
- [30] Pudenz, K. L., Albash, T. & Lidar, D. A. Quantum annealing correction for random Ising problems. *Phys. Rev. A* **91**, 042302 (2015). URL <http://link.aps.org/doi/10.1103/PhysRevA.91.042302>.
- [31] Vinci, W., Albash, T., Paz-Silva, G., Hen, I. & Lidar, D. A. Quantum annealing correction with minor embedding. *Phys. Rev. A* **92**, 042310– (2015). URL <http://link.aps.org/doi/10.1103/PhysRevA.92.042310>.
- [32] Mishra, A., Albash, T. & Lidar, D. Performance of two different quantum annealing correction codes. *arXiv:1508.02785* (2015). URL <http://arXiv.org/abs/1508.02785>.
- [33] Matsuura, S., Nishimori, H., Albash, T. & Lidar, D. A. Mean Field Analysis of Quantum Annealing Correction. *arXiv:1510.07709* (2015). URL <http://arXiv.org/abs/1510.07709>.
- [34] Aliferis, P., Gottesman, D. & Preskill, J. Quantum accuracy threshold for concatenated distance-3 codes. *Quantum Inf. Comput.* **6**, 97 (2006). URL <http://www.rintonpress.com/xqic6/qic-6-2/097-165.pdf>.
- [35] Johnson, M. W. *et al.* Quantum annealing with manufactured spins. *Nature* **473**, 194–198 (2011).
- [36] Johnson, M. W. *et al.* A scalable control system for a superconducting adiabatic quantum optimization processor. *Superconductor Science and Technology* **23**, 065004 (2010). URL <http://stacks.iop.org/0953-2048/23/i=6/a=065004>.
- [37] Berkley, A. J. *et al.* A scalable readout system for a superconducting adiabatic quantum optimization system. *Superconductor Science and Technology* **23**, 105014 (2010). URL <http://stacks.iop.org/0953-2048/23/i=10/a=105014>.
- [38] Harris, R. *et al.* Experimental investigation of an eight-qubit unit cell in a superconducting optimization processor. *Phys. Rev. B* **82**, 024511 (2010).
- [39] Boixo, S. *et al.* Evidence for quantum annealing with more than one hundred qubits. *Nat. Phys.* **10**, 218–224 (2014).
- [40] Shin, S. W., Smith, G., Smolin, J. A. & Vazirani, U. How “Quantum” is the D-Wave Machine? *arXiv:1401.7087* (2014). URL <http://arXiv.org/abs/1401.7087>.
- [41] Albash, T., Rønnow, T. F., Troyer, M. & Lidar, D. A. Reexamining classical and quantum models for the D-Wave One pro-

- cessor. *Eur. Phys. J. Spec. Top.* **224**, 111–129 (2015). URL dx.doi.org/10.1140/epjst/e2015-02346-0.
- [42] Albash, T., Vinci, W., Mishra, A., Warburton, P. A. & Lidar, D. A. Consistency tests of classical and quantum models for a quantum annealer. *Phys. Rev. A* **91**, 042314– (2015). URL <http://link.aps.org/doi/10.1103/PhysRevA.91.042314>.
- [43] Crowley, P. J. D., Durić, T., Vinci, W., Warburton, P. A. & Green, A. G. Quantum and classical dynamics in adiabatic computation. *Phys. Rev. A* **90**, 042317– (2014). URL <http://link.aps.org/doi/10.1103/PhysRevA.90.042317>.
- [44] Martin-Mayor, V. & Hen, I. Unraveling Quantum Annealers using Classical Hardness. *arXiv:1502.02494* (2015). URL <http://arXiv.org/abs/1502.02494>.
- [45] King, A. D., Lanting, T. & Harris, R. Performance of a quantum annealer on range-limited constraint satisfaction problems. *arXiv:1502.02098* (2015). URL <http://arXiv.org/abs/1502.02098>.
- [46] Vinci, W. *et al.* Hearing the shape of the Ising model with a programmable superconducting-flux annealer. *Scientific reports* **4** (2014).
- [47] Bunyk, P. I. *et al.* Architectural Considerations in the Design of a Superconducting Quantum Annealing Processor. *IEEE Transactions on Applied Superconductivity* **24**, 1–10 (Aug. 2014).
- [48] Venturelli, D. *et al.* Quantum Optimization of Fully Connected Spin Glasses. *Phys. Rev. X* **5**, 031040– (2015). URL <http://link.aps.org/doi/10.1103/PhysRevX.5.031040>.
- [49] Kato, T. On the adiabatic theorem of Quantum Mechanics. *J. Phys. Soc. Jap.* **5**, 435 (1950).
- [50] Jansen, S., Ruskai, M.-B. & Seiler, R. Bounds for the adiabatic approximation with applications to quantum computation. *J. Math. Phys.* **48**, – (2007). URL <http://scitation.aip.org/content/aip/journal/jmp/48/10/10.1063/1.2798382>.
- [51] Lidar, D. A., Rezakhani, A. T. & Hamma, A. Adiabatic approximation with exponential accuracy for many-body systems and quantum computation. *J. Math. Phys.* **50**, – (2009). URL <http://scitation.aip.org/content/aip/journal/jmp/50/10/10.1063/1.3236685>.
- [52] Wiebe, N. & Babcock, N. S. Improved error-scaling for adiabatic quantum evolutions. *New J. Phys.* **14**, 013024 (2012).
- [53] Ge, Y., Molnár, A. & Cirac, J. I. Rapid adiabatic preparation of injective PEPS and Gibbs states. *arXiv:1508.00570* (2015). URL <http://arXiv.org/abs/1508.00570>.
- [54] Avron, J. E., Fraas, M., Graf, G. M. & Grech, P. Adiabatic Theorems for Generators of Contracting Evolutions. *Comm. Math. Phys.* **314**, 163–191 (2012). URL dx.doi.org/10.1007/s00220-012-1504-1.
- [55] Venuti, L. C., Albash, T., Lidar, D. A. & Zanardi, P. Adiabaticity in open quantum systems. *arXiv:1508.05558* (2015). URL <http://arXiv.org/abs/1508.05558>.
- [56] Choi, V. Minor-embedding in adiabatic quantum computation: II. Minor-universal graph design. *Quant. Inf. Proc.* **10**, 343–353 (2011). URL dx.doi.org/10.1007/s11128-010-0200-3.
- [57] Cai, J., Macready, W. G. & Roy, A. A practical heuristic for finding graph minors. *arXiv:1406.2741* (2014). URL <http://arXiv.org/abs/1406.2741>.
- [58] Boothby, T., King, A. D. & Roy, A. Fast clique minor generation in Chimera qubit connectivity graphs. *arXiv:1507.04774* (2015). URL <http://arXiv.org/abs/1507.04774>.
- [59] Kaminsky, W. M., Lloyd, S. & Orlando, T. P. *Quantum Computing and Quantum Bits in Mesoscopic Systems*, chap. 25, 229–236 (Springer, New York, 2004).
- [60] Klymko, C., Sullivan, B. D. & Humble, T. S. Adiabatic quantum programming: minor embedding with hard faults. *Quantum Information Processing* **13**, 709–729 (2014). URL <http://dx.doi.org/10.1007/s11128-013-0683-9>.
- [61] Rønnow, T. F. *et al.* Defining and detecting quantum speedup. *Science* **345**, 420–424 (2014).
- [62] Hen, I. *et al.* Probing for quantum speedup in spin-glass problems with planted solutions. *Phys. Rev. A* **92**, 042325– (2015). URL <http://link.aps.org/doi/10.1103/PhysRevA.92.042325>.
- [63] Albash, T., Lidar, D., Matsuura, S., Nishimori, H. & Vinci, W. In preparation.
- [64] Seoane, B. & Nishimori, H. Many-body transverse interactions in the quantum annealing of the p-spin ferromagnet. *J. Phys. A* **45**, 435301 (2012). URL <http://stacks.iop.org/1751-8121/45/i=43/a=435301>.
- [65] Bray, A. J. & Moore, M. A. Replica theory of quantum spin glasses. *Journal of Physics C: Solid State Physics* **13**, L655 (1980). URL <http://stacks.iop.org/0022-3719/13/i=24/a=005>.
- [66] Krzakala, F., Rosso, A., Semerjian, G. & Zamponi, F. Path-integral representation for quantum spin models: Application to the quantum cavity method and Monte Carlo simulations. *Phys. Rev. B* **78**, 134428 (2008). URL <http://link.aps.org/doi/10.1103/PhysRevB.78.134428>.
- [67] Choi, V. Minor-embedding in adiabatic quantum computation: I. The parameter setting problem. *Quant. Inf. Proc.* **7**, 193–209 (2008). URL dx.doi.org/10.1007/s11128-008-0082-9.
- [68] Kirkpatrick, S. & Sherrington, D. Infinite-ranged models of spin-glasses. *Physical Review B* **17**, 4384–4403 (1978). URL <http://link.aps.org/doi/10.1103/PhysRevB.17.4384>.
- [69] Katzgraber, H. G., Hamze, F. & Andrist, R. S. Glassy Chimeras Could Be Blind to Quantum Speedup: Designing Better Benchmarks for Quantum Annealing Machines. *Phys. Rev. X* **4**, 021008– (2014). URL <http://link.aps.org/doi/10.1103/PhysRevX.4.021008>.
- [70] Martoňák, R., Santoro, G. E. & Tosatti, E. Quantum annealing by the path-integral Monte Carlo method: The two-dimensional random Ising model. *Phys. Rev. B* **66**, 094203 (2002).
- [71] Santoro, G. E., Martoňák, R., Tosatti, E. & Car, R. Theory of Quantum Annealing of an Ising Spin Glass. *Science* **295**, 2427–2430 (2002).
- [72] Boixo, S. *et al.* Computational Role of Collective Tunneling in a Quantum Annealer. *arXiv:1411.4036* (2014). URL <http://arXiv.org/abs/1411.4036>.
- [73] Wolff, U. Collective Monte Carlo Updating for Spin Systems. *Phys. Rev. Lett.* **62**, 361–364 (1989). URL <http://link.aps.org/doi/10.1103/PhysRevLett.62.361>.

Appendix A: Experimental Methods

We tested NQAC on the DW2 quantum annealing device at the University of Southern California’s Information Sciences Institute (USC-ISI), which has been described in numerous previous publications (e.g., see [42]). The largest complete graph that can be embedded on this device, featuring 504 active qubits, is a K_{32} .

We determined an experimental value of the success probability $P_C(\alpha, \gamma)$ as a function of the energy penalty strength γ . All figures show, whenever the γ dependence is not explicitly considered, the optimal value $P_C(\alpha) = \max_{\gamma} P_C(\alpha, \gamma)$, with $\gamma \in \{0.05, 0.1, 0.2, \dots, 0.9, 1\}$. We used the same penalty value for both the nesting and the ME. In principle these two values can be optimized separately for improved performance, but we did not pursue this here, since the resulting improvement is small, as shown in Fig. 4, and costly since each instance needs to be rerun at all penalty settings.

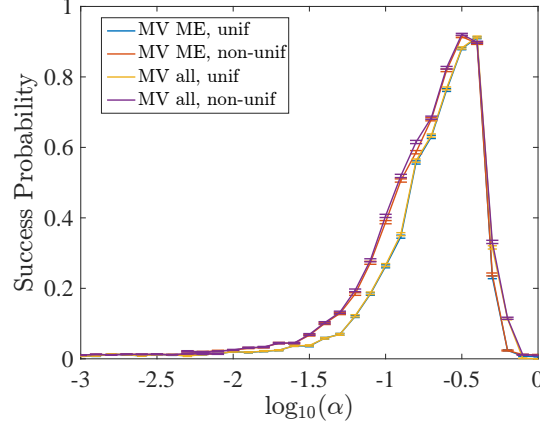


FIG. 4. Effect of separately optimizing γ for ME and penalties. The plot shows the success probability from SQA simulations, for NQAC applied to a random antiferromagnetic K_8 with 10,000 sweeps, $\sigma = 0.05$ noise, Choi embedding, with $\beta = 0.1$. The results obtained after separately optimizing the penalty for the nesting and for the ME are denoted “non-unif”, while the results for using a single penalty for both (the strategy used in the main text) is denoted “unif”. The former results in a small improvement. Also shown is that separate (“MV ME”) or joint (“MV all”) majority vote decoding of the nesting and the ME has no effect.

Each $P_C(\alpha, \gamma)$ is the overall success probability after 2×10^4 annealing runs obtained by implementing 20 programming cycles of 10^3 runs each. A sufficiently large number of programming cycles is necessary to average out intrinsic control errors (ICE) that, as explained in the main text, prevent the physical couplings to be set with a precision better than $\sim 5\%$. To further remove possible sources of systematic noise, at each programming cycle we perform a random gauge transformation on the values of the physical qubits. A permutation of the $C \times N$ vertices is a symmetry of the nested graph but it is not a symmetry of the encoded Hamiltonian obtained after ME. This is because the $C \times N$ chains of physical qubits are physically distinguishable. In each programming cycle we also then performed a random permutation of the vertices of the nested graph, before proceeding to the ME. Error bars correspond to the standard error of the mean of the 20 $P_C(\alpha)$ values.

Appendix B: Mean Field Analysis of the Partition Function

In this section we sketch how to compute the partition function of the logical problem [Eq. (3) of the main text], in order to analyze the effect of nesting. Full details will be given in a subsequent publication [63]. In the main text we were concerned with Hamiltonians of the form

$$H = B(t)(H^x + H^z) \quad (\text{B1})$$

where

$$H^x = [A(t)/B(t)]H_X = -\Gamma(t) \sum_{i=1}^N \sum_{c_i=1}^C \sigma_{ic_i}^x \quad (\text{B2a})$$

$$H^z = \bar{H}_P = \sum_{i,j=1}^N \sum_{c_i,c'_j=1}^C J_{(ic_i),(jc'_j)} \sigma_{ic_i}^z \sigma_{jc'_j}^z \quad (\text{B2b})$$

$$= \frac{J}{N} \sum_{i \neq j} \sum_{c_i,c'_j=1}^C \sigma_{ic_i}^z \sigma_{jc'_j}^z - \gamma \sum_{i=1}^N \sum_{c_i \neq c'_i} \sigma_{ic_i}^z \sigma_{ic'_i}^z, \quad (\text{B2c})$$

$A(t), B(t)$ have dimensions of energy, and where J and γ are dimensionless, and have each absorbed a factor of $1/2$ to account for double counting. Note that both H^x and H^z are extensive (proportional to N). Throughout we use $\sigma_{ic}^z \equiv \sigma_{ic_i}^z$ ($\sigma_{ic}^x \equiv \sigma_{ic_i}^x$) to denote the Pauli z (x) operator acting on physical qubit c of encoded qubit i .

We define the collective variables

$$S_i^x \equiv \frac{1}{C} \sum_{c_i=1}^C \sigma_{ic_i}^x, \quad S_i^z \equiv \frac{1}{C} \sum_{c_i=1}^C \sigma_{ic_i}^z, \quad S^x \equiv \frac{1}{N} \sum_{i=1}^N S_i^x, \quad S^z \equiv \frac{1}{N} \sum_{i=1}^N S_i^z. \quad (\text{B3})$$

We can interpret S_i^x and S_i^z as the mean transverse and longitudinal fields on logical qubit i , respectively. Then

$$H^x = -\Gamma(t)C \sum_{i=1}^N S_i^x = -NCT(t)S^x, \quad (\text{B4})$$

and

$$\bar{H}_P = \frac{J}{N}C^2 \sum_{i,j} S_i^z S_j^z - \left(\frac{J}{N} + \gamma \right) \sum_{i=1}^N \sum_{c_i,c'_i} \sigma_{ic_i}^z \sigma_{ic'_i}^z + \gamma \sum_{i=1}^N \sum_{c_i} (\sigma_{ic_i}^z)^2, \quad (\text{B5})$$

but the last term is a constant [equal to $\gamma NC \mathbb{1}$], so it can be ignored. Therefore, up to a constant we have

$$\bar{H}_P = JNC^2 \left((S^z)^2 - \lambda \frac{1}{N} \sum_{i=1}^N (S_i^z)^2 \right), \quad (\text{B6})$$

where

$$\lambda = \frac{\gamma}{J} + \frac{1}{N} \geq 0, \quad (\text{B7})$$

encodes the penalty strength; the $1/N$ correction will disappear in the thermodynamic limit. Note that $\frac{1}{N} \sum_{i=1}^N (S_i^z)^2 = O(1)$, so that $\lambda \frac{1}{N} \sum_{i=1}^N (S_i^z)^2 = O(1)$, like $(S^z)^2$, and hence \bar{H}_P is extensive in N , as it should be.

The form (B6) for \bar{H}_P shows that the NQAC Hamiltonian in the fully antiferromagnetic $K_{N \times C}$ case can be interpreted as describing the collective evolution of all logical qubits. The term $\lambda \sum_{i=1}^N (S_i^z)^2$ favors all the spins of each logical qubit (where by spin we mean the qubit at $t = t_f$) being aligned, since this maximizes each summand.

1. Partition Function Calculation

We are interested in the partition function

$$Z = \text{Tr} e^{-\beta H} = \text{Tr} e^{-\beta B(t)[H^x + H^z]} = \text{Tr} e^{-\theta[H^x + H^z]}, \quad (\text{B8})$$

where $\theta = \beta B(t)$ is the dimensionless inverse temperature. We write the partition function explicitly as [64]

$$Z = \sum_{\{\sigma^z\}} \langle \{\sigma^z\} | \exp[-\theta(H^z + H^x)] | \{\sigma^z\} \rangle = \lim_{M \rightarrow \infty} Z_M, \quad (\text{B9})$$

where $\sum_{\{\sigma^z\}}$ is a sum over all possible 2^{CN} spin configurations in the z basis, and $|\{\sigma^z\}\rangle = \otimes_{i=1}^N \otimes_{c=1}^C |\sigma_{ic}^z\rangle$. Z_M is determined using the Trotter-Suzuki formula $e^{A+B} = \lim_{M \rightarrow \infty} (e^{A/M} e^{B/M})^M$:

$$Z_M = \sum_{\{\sigma^z\}} \langle \{\sigma^z\} | \left(\exp \left[-\frac{\theta}{M} H^z \right] \exp \left[-\frac{\theta}{M} H^x \right] \right)^M | \{\sigma^z\} \rangle. \quad (\text{B10})$$

After a lengthy calculation [63] we find

$$Z \approx \int \Pi_j \mathcal{D}m_j \mathcal{D}\tilde{m}_j e^{N \left[\frac{1}{N} \sum_{j=1}^N \left\{ C \ln \left[2 \cosh \left((\theta\Gamma)^2 - (\tilde{m}_j/C)^2 \right)^{1/2} \right] + m_j (i\tilde{m}_j + \theta J C^2 \lambda m_j) \right\} - \theta J C^2 \langle m \rangle^2 \right]}, \quad (\text{B11})$$

where $\langle m \rangle \equiv \frac{1}{N} \sum_{j=1}^N m_j$, and where m_j is the Hubbard-Stratonovich field that represents $S_j^z(\alpha)$ after the static approximation (i.e., dropping the α dependence) [65, 66]. The second Hubbard-Stratonovich field \tilde{m}_j acts as a Lagrange multiplier.

2. Free energy

In the large β (low temperature) limit, the partition function is dominated by the global minimum. This minimum is given by $\langle m \rangle = 0$, which corresponds to either a paramagnetic phase (all $m_j = 0$) or a symmetric phase ($m_j = \pm m$ in equal numbers). It can be shown that the system undergoes a second order QPT, with the critical point moving to the left as C and γ grow [63]. Using a saddle point analysis of the partition function we can show that $\tilde{m}_j = \pm 2i\theta C^2 J \lambda m$, and hence, the dominant contribution to the partition function is given by:

$$Z_C \approx e^{N \left\{ C \ln \left[2 \cosh \left((\theta\Gamma)^2 + (2\theta J C \lambda m)^2 \right)^{1/2} \right] - \theta J C^2 \lambda m^2 \right\}} \quad (\text{B12a})$$

$$= e^{N \left\{ C \ln \left[2 \cosh \left([\beta A(t)]^2 + [2\beta B(t) C (\gamma + \frac{J}{N}) m]^2 \right)^{1/2} \right] - \beta B(t) C^2 (\gamma + \frac{J}{N}) m^2 \right\}}. \quad (\text{B12b})$$

For $B(t) > 0$ and in the low temperature limit ($\theta \gg 1$) we can approximate $2 \cosh(\theta|x|)$ as $e^{\theta|x|}$,

$$Z_C \approx e^{N\theta \left\{ ((C\Gamma)^2 + (2J\lambda C^2 m)^2)^{1/2} - J\lambda C^2 m^2 \right\}} \quad (\text{B13a})$$

$$= e^{N\beta \left\{ [CA(t)]^2 + [2B(t)(\gamma + \frac{J}{N}) C^2 m]^2 \right\}^{1/2} - B(t)(\gamma + \frac{J}{N}) C^2 m^2} = e^{-\beta N F}, \quad (\text{B13b})$$

where in the second line we reintroduced the physical inverse temperature β [recall Eq. (B8)]. Factoring out C^2 and taking the large N limit then directly yields the free energy density expression (5) given in the main text.

Appendix C: Additional Numerical Data

Figure 5(a) shows that the saturation of μ_C at large C is removed when the number of sweeps is increased. The thermal state, where the system has fully thermalized, can be understood as the limit of an infinite number of sweeps. Figure 5(b) shows that the saturation is fully removed for the thermal state (generated using parallel tempering), and nesting is then equivalent to an energy (or temperature) boost close to the ideal result $\mu_C^{\max} = C^2$. This suggests that for a sufficiently large sweep number, performance can be brought to near the ideal result.

Figure 6 gives further evidence that nesting can be interpreted as an effective reduction of temperature by studying the success probability associated with the thermal distribution on the ME. We used parallel tempering (PT) to sample from the thermal state associated with the ME of the different NQAC cases shown in Fig. 2, and decoded using majority voting. We find that the thermal state at different temperatures but fixed C , exhibits the same qualitative behavior as the thermal state at fixed temperature but different C [see Fig. 6(a) vs. Fig. 6(b)]. Therefore, the performance improvement associated with reducing the temperature can also be reproduced by increasing C . This enforces that the energy boost can also be interpreted as decrease of the effective temperature of the device. We also find that the thermal state exhibits an energy boost scaling of $\mu_C \sim C^2$ [see Fig. 6(c)].

Appendix D: Choi and Heuristic Embeddings

The ‘‘Chimera’’ hardware connectivity graph of the D-Wave devices allows for a ME of complete graphs as described in Refs. [56, 67]. In the main text we called this the ‘‘Choi minor embedding’’. The Choi ME of a K_{32} graph is shown in Fig. 7(a).

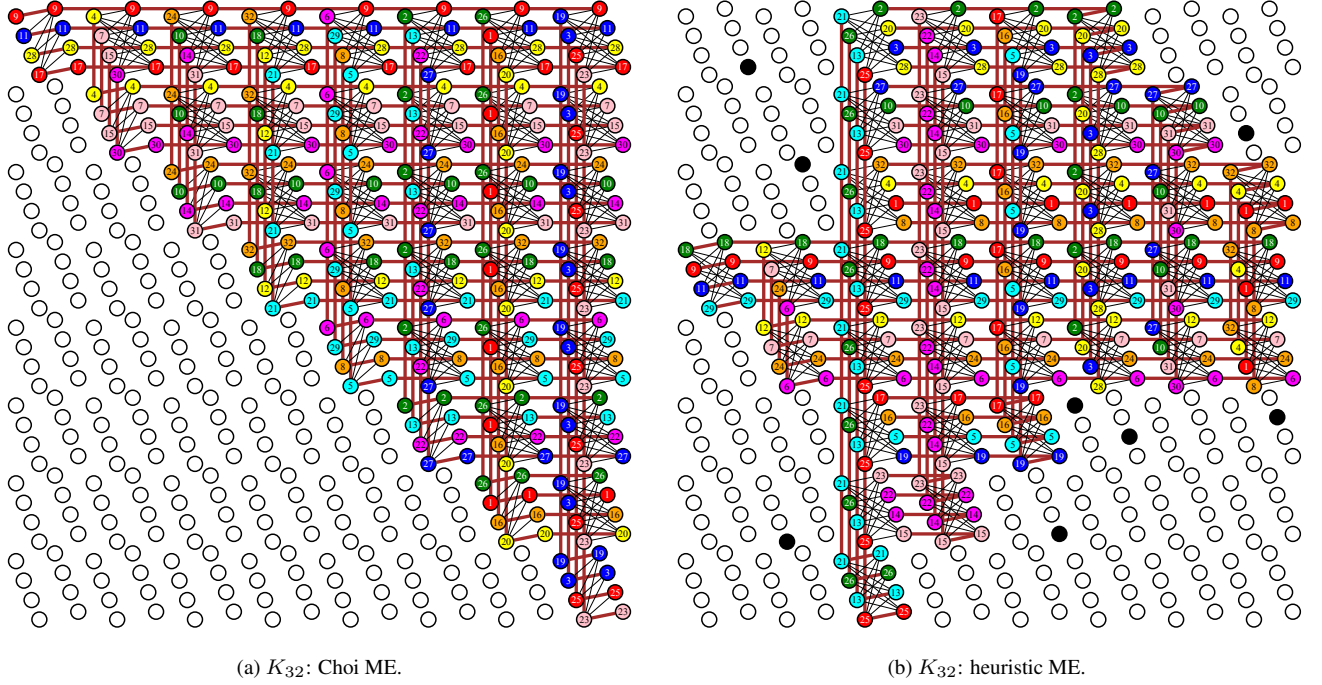


FIG. 7. MEs of a K_{32} . We used these, e.g., to minor-embed a $C = 8$ nesting of a K_4 , or a $C = 4$ nesting of a K_8 . (a) The Choi embedding implemented on a perfect Chimera graph. (b) A heuristic ME for the actual DW2 device used in this work, whose Chimera graph contains 8 unusable qubits (black circles). Different colors (and labels) denote chains representing minor-embedded logical qubits. Black (thin) lines are logical couplings, while brown (thick) lines represent energy penalties (ferromagnetic couplings).

Figure 9 displays results for NQAC applied to an “easier- K_{10} ”:

$$K_{10}^e = \begin{pmatrix} 0 & 0.2 & 0.7 & 0.8 & 0.5 & 0.3 & 0.8 & 0.9 & 0.4 & 0.1 \\ 0 & 0 & 0.1 & 0.1 & 0.4 & 0.7 & 0.3 & 0.3 & 0.9 & 0.1 \\ 0 & 0 & 0 & 0.3 & 0.8 & 0.7 & 0.6 & 0.9 & 0.6 & 0.6 \\ 0 & 0 & 0 & 0 & 0.8 & 0.2 & 0.7 & 0.3 & 0.6 & 0.8 \\ 0 & 0 & 0 & 0 & 0 & 0.2 & 0.9 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0.4 & 0.3 & 0.2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0.8 & 0.6 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.8 & 0.5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad (E3)$$

and a “harder- K_{10} ”:

$$K_{10}^h = \begin{pmatrix} 0 & 0.6 & 0.9 & 0.8 & 0.5 & 1 & 0.4 & 0.2 & 0.1 & 0.5 \\ 0 & 0 & 0.8 & 0.9 & 0.1 & 0.6 & 0.2 & 0.7 & 0.7 & 0.9 \\ 0 & 0 & 0 & 0.8 & 0.6 & 0.3 & 0.8 & 0.2 & 0.6 & 0.6 \\ 0 & 0 & 0 & 0 & 0.1 & 0.3 & 0.8 & 0.4 & 0.6 & 0.5 \\ 0 & 0 & 0 & 0 & 0 & 0.7 & 0.6 & 0.4 & 0.3 & 0.1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.1 & 1 & 0.9 & 0.6 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.9 & 0.9 & 0.9 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1 & 1.0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \quad (E4)$$

In all cases we display results up to nesting level $C = 3$.

Figure 10 shows the optimal penalty strength as a function of the energy scale α for the four instances considered. A saturation of the optimal penalty is visible at the maximal possible value $|\gamma| = 1$ for α close to 1, implying that the true optimal penalty values are > 1 in this range.

Figure 11 shows that the antiferromagnetic harder- K_8 problem considered in the main text, as well as the easier- K_8 problem, also admit a data collapse (left), to the left of the peak. Recall that the peak is due to having reached the maximum penalty value, as illustrated in Fig. 10. The associated scaling of the energy boost μ_C is shown in the right column, yielding $\mu_C \sim C^{1.32}$ (harder- K_8) and $\mu_C \sim C^{1.26}$ (easier- K_8). Figure 12 shows the same for harder- K_{10} and easier- K_{10} problems. There we find $\mu_C \sim C^{1.34}$ for both problems.

Appendix F: Determination of μ_C

To determine the values of μ_C and estimate error bars, we proceeded as follows. First, we used smoothing splines to determine a continuous interpolation $P_C^{\text{mid}}(\alpha)$ of the discrete data points $P_C(\alpha)$. In the same way we also determined the higher and lower interpolating curves $P_C^{\text{high}}(\alpha)$ and $P_C^{\text{low}}(\alpha)$ for the data points $P_C(\alpha) + \delta P_C(\alpha)$ and $P_C(\alpha) - \delta P_C(\alpha)$ respectively, where $\delta P_C(\alpha)$ denotes the standard error of $P_C(\alpha)$. A reference value α_C^{mid} was then determined such that $P_C^{\text{mid}}(\alpha_C^{\text{mid}}) = P_0$, where we used the smooth interpolation of the experimental data. The energy boost was then determined as $\mu_C = \alpha_1^{\text{mid}} / \alpha_C^{\text{mid}}$. P_0 is an arbitrarily chosen reference value where the different $P_C(\alpha)$ curves are overlapped. This reference serves as a base point for computing μ_C . As shown in the main text for the K_4 , the overlap of the P_C data over the entire α range means that the specific choice of P_0 is arbitrary.

We similarly determined $\mu_C^{\text{high}} = \alpha_1^{\text{high}} / \alpha_C^{\text{high}}$ and $\mu_C^{\text{low}} = \alpha_1^{\text{low}} / \alpha_C^{\text{low}}$ using the corresponding interpolating curves. The error bars shown in the figures were then centered at μ_C , with lower and upper error bars being μ_C^{high} and μ_C^{low} , respectively.

Appendix G: Numerical Methods

We reported results based on quantum Monte Carlo techniques in the main text. Here we briefly review this technique. Simulated Quantum Annealing (SQA) is a quantum Monte Carlo based algorithm whereby Monte Carlo dynamics are used to sample from the instantaneous Gibbs state associated with the Hamiltonian $H(t)$ of the system. The state at the end of the quantum Monte Carlo simulation of the quantum Hamiltonian $H(t)$ is used as the initial state for the next Monte Carlo simulation with Hamiltonian $H(t + \Delta t)$. This is repeated until $H(t_f)$ is reached. SQA was originally proposed as an optimization algorithm [70, 71], but it has since gained traction as a computationally efficient classical description for $T > 0$ quantum annealers [39, 41, 61, 62]. An important caveat is that SQA does not capture the unitary dynamics of the quantum system, but it is hoped that the sampling of the instantaneous Gibbs state captures thermal processes in the quantum annealer, which may be the dominant dynamics if the evolution is sufficiently slow. Although there is strong evidence that SQA does not completely capture the final-time output of the D-Wave processors [41, 72], at present it is the only viable means to simulate large ($\gtrsim 15$ qubits) open QA systems. We used discrete-time quantum Monte Carlo in our simulations with the number of Trotter slices fixed to 64. Spin updates were performed via Wolff-cluster updates [73] along the Trotter direction only.

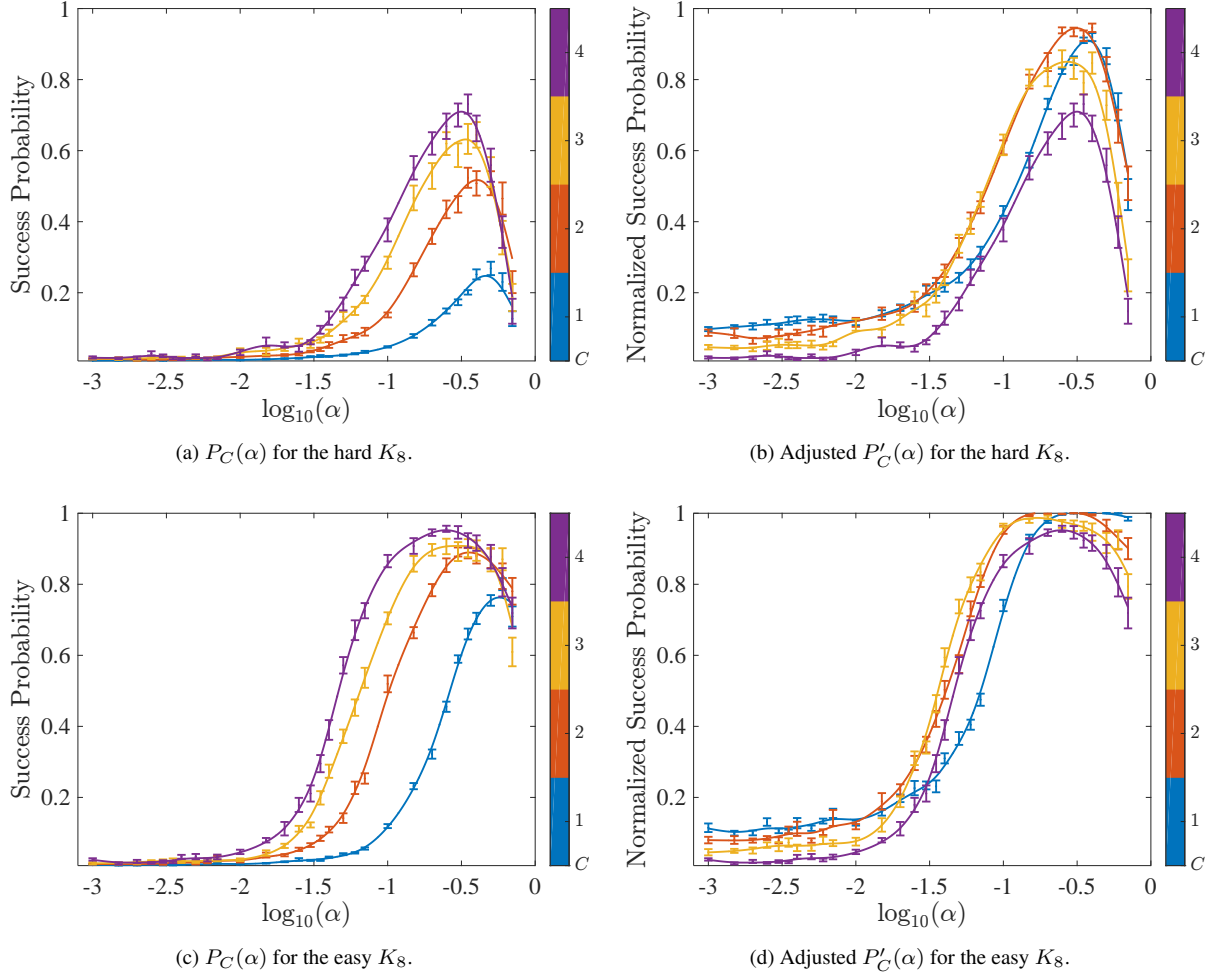


FIG. 8. Random antiferromagnetic K_8 nesting: experimental results for a harder and easier instance.

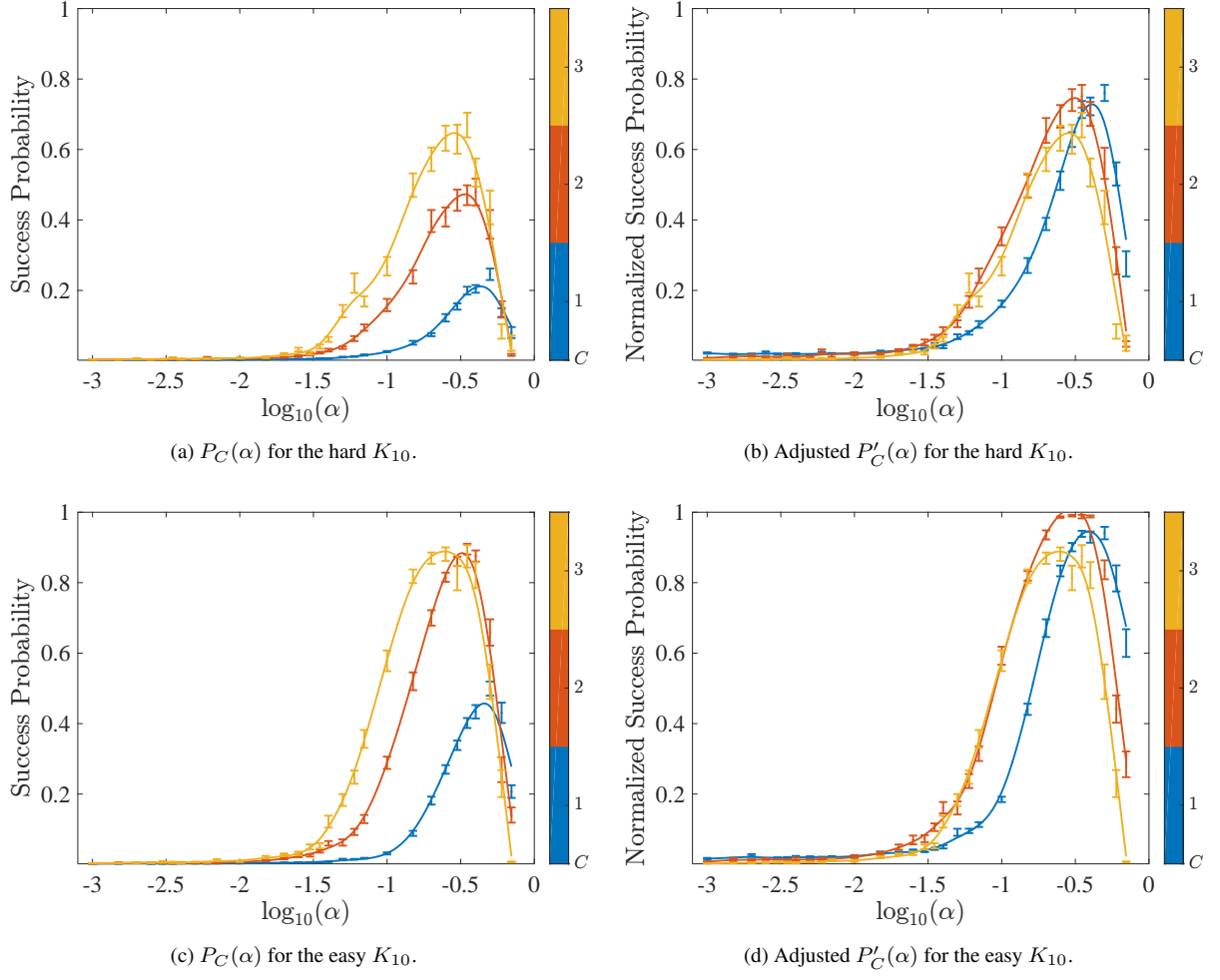
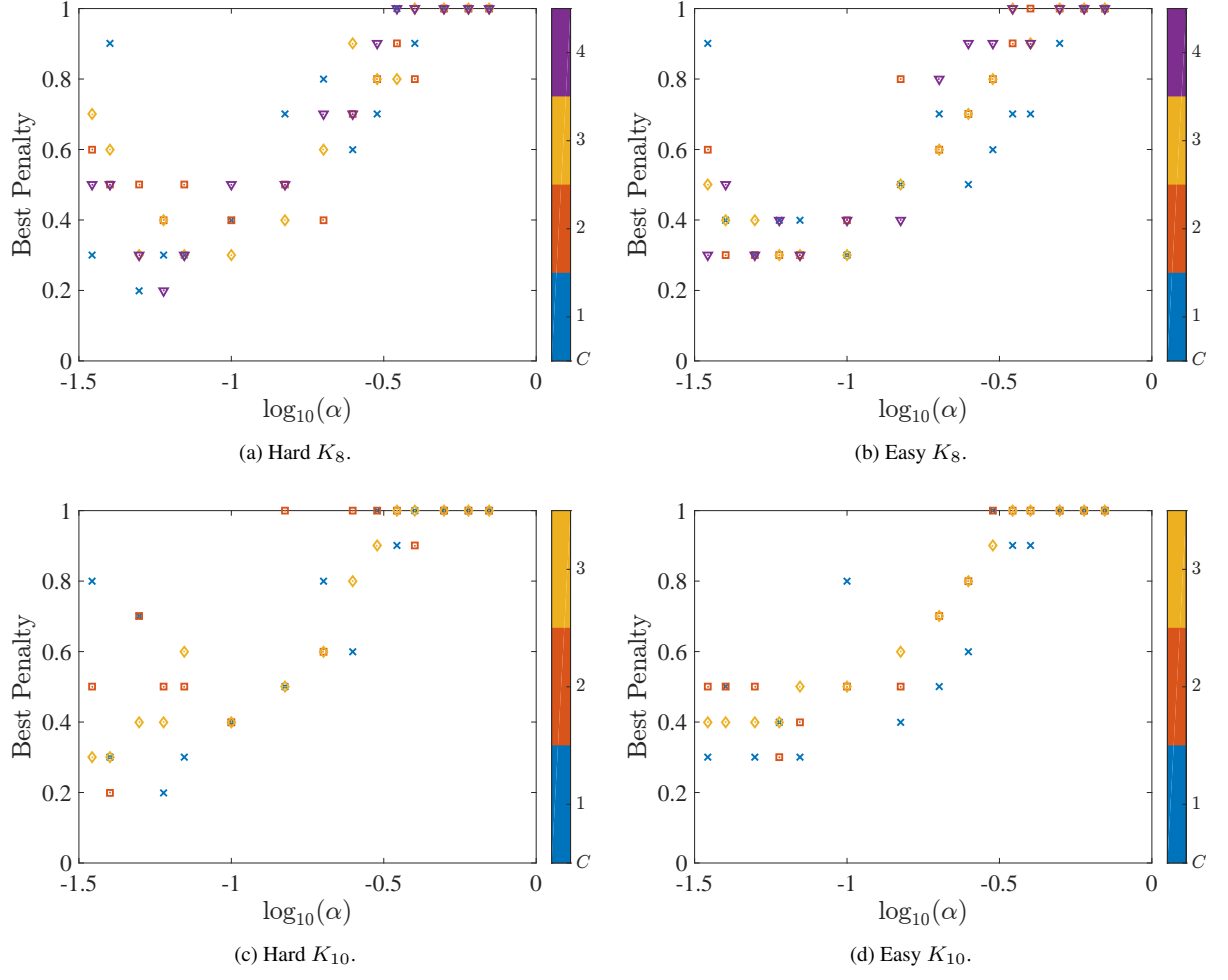


FIG. 9. Random antiferromagnetic K_{10} nesting: experimental results for a harder and easier instance.

FIG. 10. Optimal penalties γ .

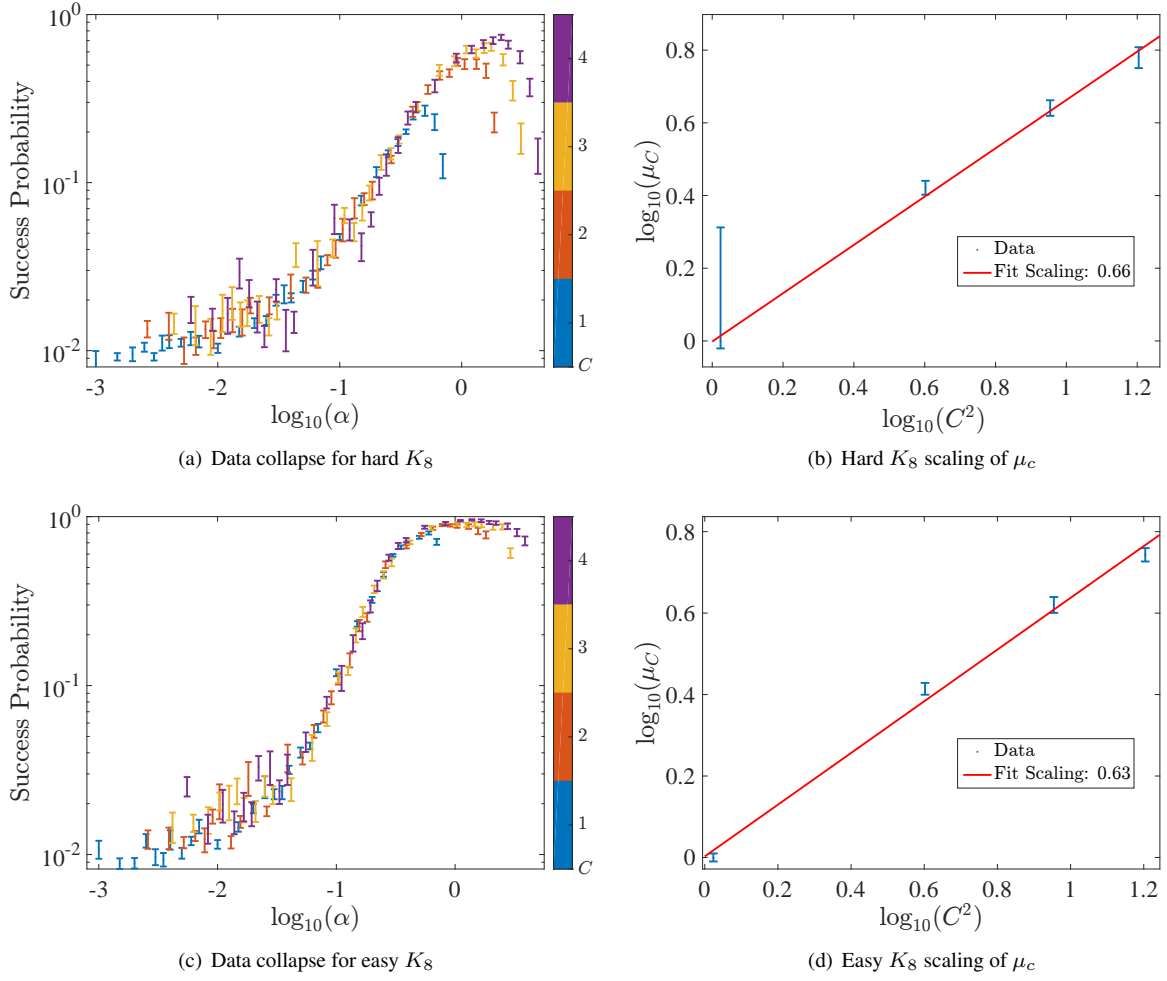
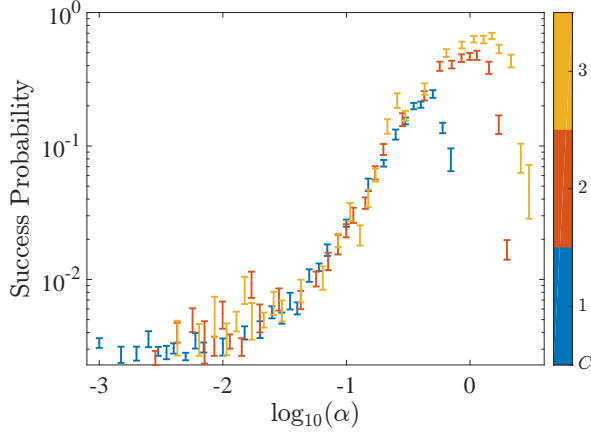
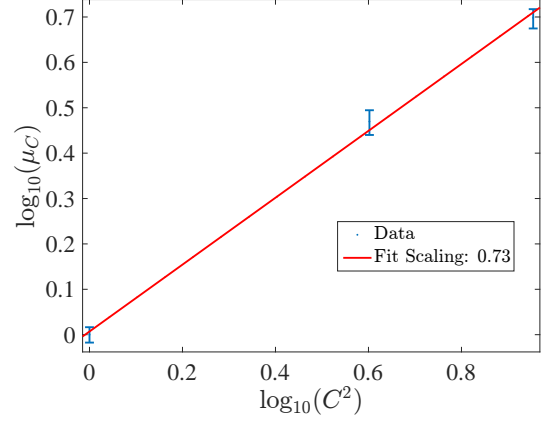
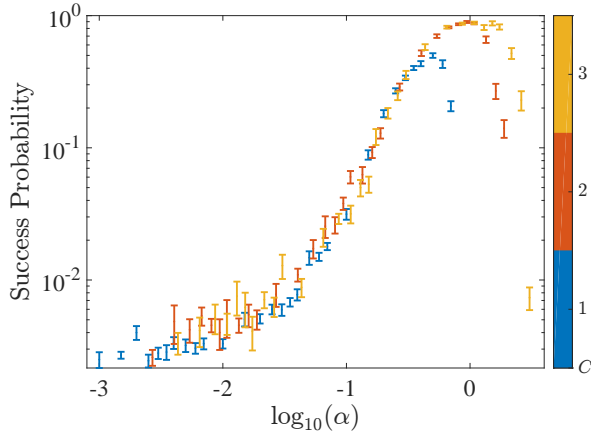
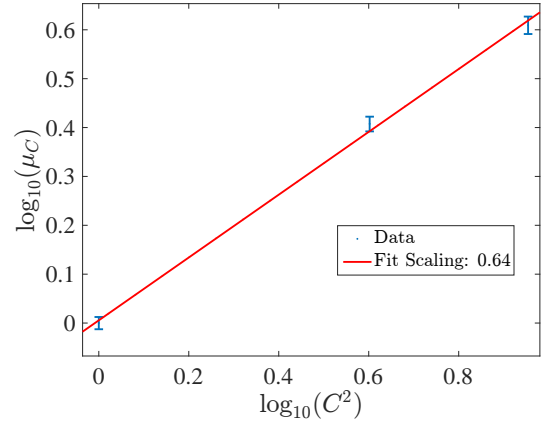


FIG. 11. Data collapse (left) and scaling of μ_C (right) for fully antiferromagnetic K_8 .

(a) Data collapse for hard K_{10} (b) Hard K_{10} scaling of μ_c (c) Data collapse for easy K_{10} (d) Easy K_{10} scaling of μ_c FIG. 12. Data collapse (left) and scaling of μ_c (right) for fully antiferromagnetic K_{10} .